

本周周报

解聪

一、企业数据进行分析

企业数据包含了很多比较专业的属性：（以下指示其中一部分）

流动资产合计，流动资产年平均余额，长期投资，固定资产合计，固定资产原价，累计折旧，固定资产净值年平均，余额，无形资产，资产总计，流动负债合计，长期负债合计，负债合计，所有者权益合计，主营业务收入，主营业务成本，主营业务税金及附加，其他业务收入，其他业务利润，营业费用，管理费用，财务费用，营业利润，投资收益，补贴收入，营业外收入，利润总额，亏损总额，利税总额，应交所得税，广告费，研究开发费，劳动、失业保险费，养老保险和医疗保险费，住房公积金和住房补贴，本年应付工资总额(贷方累计发生额)，本年应付福利费总额(贷方累计发生额),本年应交增值税,本年进项税额,本年销项税额,工业中间投入合计

很多听都没听过的术语。为了简化便于分析，简单选取了 8 个我自己了解并认为能够反应企业行为的维度：员工数，总资产，负债，年收入，利润，亏损，广告投入和科研投入。

使用 Google Refine 先简单处理了数据

▼ company_name	▼ employer_num	▼ year_income	▼ total_assets	▼ debt	▼ total_profits	▼ deficit	▼ advertisement	▼ research
北京市燕山水泥厂	450	139288	183754	138756	4964	0	0	2530
北京衬衫厂	319	17008	74707	41625	-3028	-3028	0	0
北京日用化学二厂	129	14540	258184	39751	-4963	-4963	0	0
北京建筑工业印刷厂	144	11932	40220	39738	-4447	-4447	0	0
北京三五零一服装厂	137	41406	132498	77974	5763	0	0	0
北京京冠毛巾有限责任公司	348	42614	82824	57145	-4421	-4421	0	0
文物出版社印刷厂	130	14775	44594	7864	-489	-489	0	0
化学工业出版社印刷厂	243	21913	88730	66409	-2831	-2831	0	0
北京外文印刷厂	545	78728	115377	21918	-1932	-1932	0	0
北京长亿参饮料有限公司	245	24534	136412	6828	1680	0	0	6
丰台桥梁厂	319	65649	241416	133853	3444	0	14	0
北京旭鑫印刷厂	319	5472	21618	44275	-2672	-2672	2	0
北京新立机械厂（国营第六九九厂）	1411	464727	791549	674483	17606	0	10	876
北京卷烟厂	795	1685781	1378928	76980	163628	0	5467	1122
中青印刷厂	497	41849	92258	17098	2	0	1	0
北京宇翔电子有限公司	370	32701	74236	56966	-10630	-10630	0	0
北京广播电影电视设备制造厂	96	62805	122344	76939	-1106	-1106	0	0
北京起重机器厂	764	86452	192998	248571	-27218	-27218	0	887
北京白菊电器集团	417	71554	409822	304215	407	0	6	105
北京远通制管有限公司	576	92103	260310	198871	4	0	0	0
北京市科通电子继电器总厂	232	18658	30089	25493	201	0	96	3165
北京瑞利分析仪器公司	400	55880	169201	54797	1931	0	142	6545
北京正东电子动力集团有限公司	658	107123	1250715	771327	3115	0	0	0
北京市德安汽车修理厂	271	18801	39794	18591	506	0	0	0

首先先观察下数据的分布。筛选掉一些本身就不太正常的数据以及一些值非常极端的数据，总共约 270,000 家企业。

员工数：大部分企业分布在 0~1000 人间

年收入：大部分企业在 0~1000 万元间

总资产：大部分企业在 0~100 万元间

负债：大部分企业在 0~100 万元间

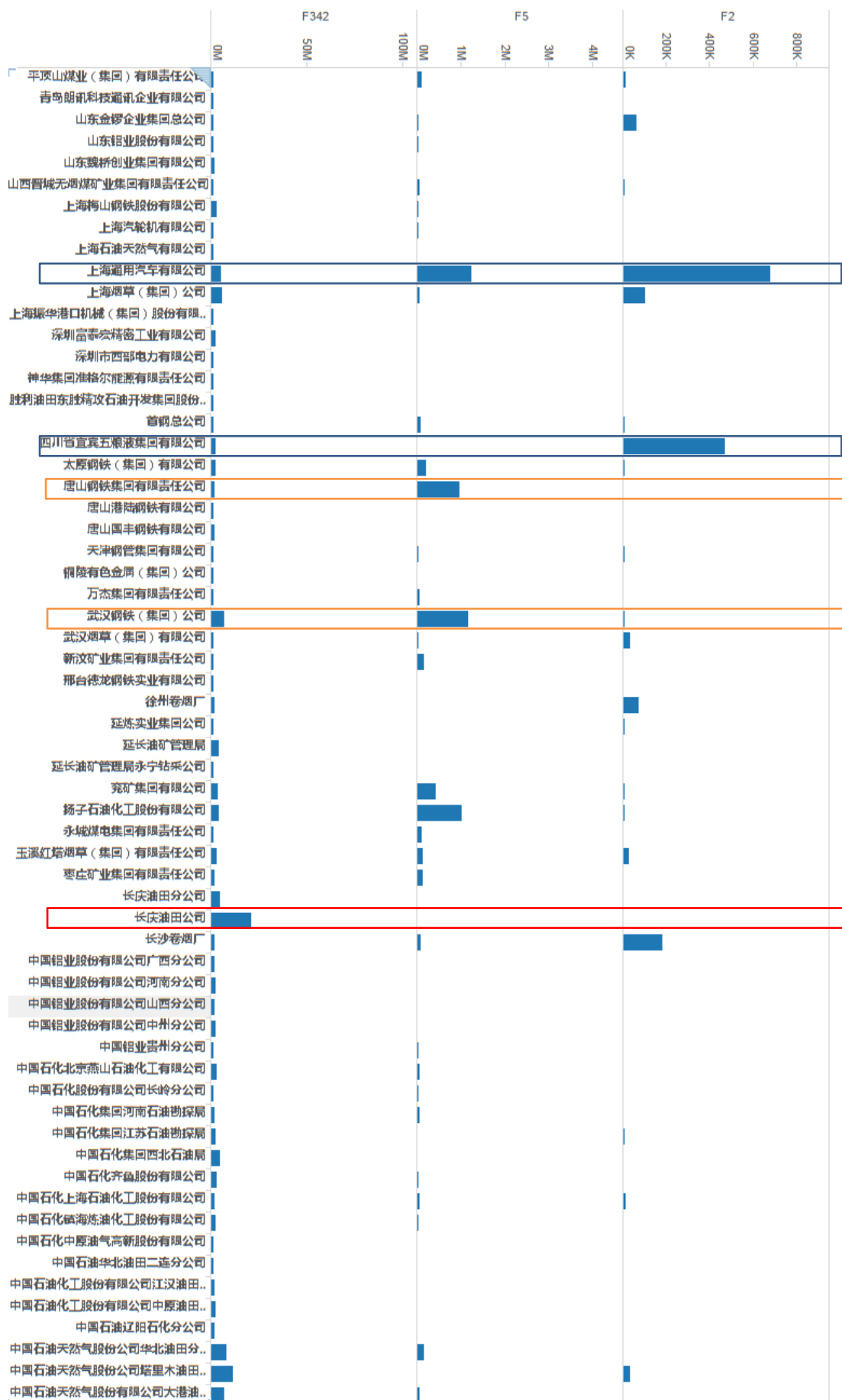
总利润与亏损：大部分企业在-10~10 万元间

广告费与科研费：大部分企业在-10~10 万元间

其次,对这些数据进行基本的可视化探索，使用了 Tablaeu

主要对于利润，广告投入以及科研投入进行分析：

下图只截取部分有盈利的公司的部分结果：（F342 是利润，F5 是科研，F2 是广告）



图中可以看出：

1. 红色框出来的公司很少投入广告，科研费用。但是其盈利是最大的。
典型代表：大庆石油，长庆油田。（主要是能源部门，反正能源紧张，这些公司不需要做太多广告，而且可以看出也不需要什么研发。）
2. 蓝色框中的公司主要是投入在广告上，基本广告费之处大于其利润。
典型代表：安利，海尔，通用，五粮液。（基本是国内或国际最著名的品牌，至少经常在电视上看到，说明其运营模式很商业化。其中食品保健品等行业的科研投入与广告投入比几乎可以忽略不计。）
3. 橙色框中是主要子集投入在研发中，甚至高过其利润。
典型代表：华为，唐山钢铁，武汉钢铁。（华为是意料之中。至于钢铁企业，原本我以为和石油企业一样，但是从数据上看，虽然都是国有垄断企业，但是这两类公司差别很大。）

下一步：

查看其他属性的经济学意义，如所有者权益合计等。

二、CCA 对商品数据的分析

这部分尚平平应该会有介绍

三、周三的论文报告” Explainers: Expert Explorations with Crafted Projections”

下周工作

1. 对企业数据进行深入分析，并且包含代码实现一部分功能，如可视化等。
2. 完成 CCA 对商品分析，看效果如何。